

Layered Logic Classifiers: Exploring the ‘And’ and ‘Or’ Relations

Zhuowen Tu¹, Piotr Dollar², and Yingnian Wu³

¹Dept. of Cognitive Science, UCSD, ²Microsoft Research, ³Department of Statistics, UCLA

ztu@ucsd.edu, pdollar@gmail.com, ywu@stat.ucla.edu

Abstract

Designing effective and efficient classifier for pattern analysis is a key problem in machine learning and computer vision. Many the solutions to the problem require to perform logic operations such as ‘and’, ‘or’, and ‘not’. Classification and regression tree (CART) include these operations explicitly. Other methods such as neural networks, SVM, and boosting learn/compute a weighted sum on features (weak classifiers), which weakly perform the ‘and’ and ‘or’ operations. However, it is hard for these classifiers to deal with the ‘xor’ pattern directly. In this paper, we propose layered logic classifiers for patterns of complicated distributions by combining the ‘and’, ‘or’, and ‘not’ operations. The proposed algorithm is very general and easy to implement. We test the classifiers on several typical datasets from the Irvine repository and two challenging vision applications, object segmentation and pedestrian detection. We observe significant improvements on all the datasets over the widely used decision stump based AdaBoost algorithm. The resulting classifiers have much less training complexity than decision tree based AdaBoost, and can be applied in a wide range of domains.

1. Introduction

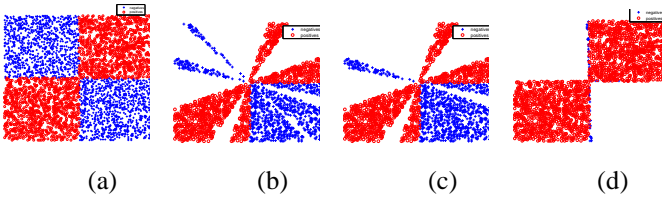


Figure 1. The xor problem. (a) shows the positive and negative points. (b) are the points classified as positives by the AdaBoost algorithm using 100 decision stump weak classifiers. (c) shows points classified as positives by the Ada-Ada algorithm which will be discussed later. (d) shows the positive points classified by the Ada-Or classifier using 10 OrBoost weak classifiers.

Classification algorithms such as decision tree [16, 2],

neural networks [1], support vector machine (SVM) have been widely used in many areas. The classification procedure in many of these algorithms can be understood as performing reasoning using logic operators (and, or, not), with deterministic or probabilistic formulations. The recent development of the AdaBoost algorithm [12] has particularly advanced the performance of many applications in the field. We focus on the AdaBoost algorithm in this paper (also called boosting together with its variations [5, 11]).

Boosting algorithms have many advantages over the traditional classification algorithms. Its asymptotical behavior when combining a large number of weak classifiers is less prone to the overfitting problem. Once trained, a boosting algorithm performs weighted sum on the selected weak classifiers. This linear summation weakly performs the ‘and’ and ‘or’ operations. In the discrete case, as long as the overall score is above the threshold, a pattern is considered as positive. This may include a combinatory combinations of the conditions. Some weak classifiers may require to be satisfied together (‘and’), and some may not as long a subset answer yes (‘or’).

In the literature, decision stump has been widely used as weak classifier due to its speed and small complexity. However, decision stump does not have strong discrimination power. A comprehensive empirical study for a wide variety of classifiers including SVM, Boosting (using decision-tree and decision-stump), neural networks, and nearest neighborhood, was reported in [6]. Each decision stump corresponds to a thresholded feature (\geq is changeable to $<$):

$$h(F_j(x), tr) = \begin{cases} +1 & \text{if } F_j(x) \geq tr \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

We call stump based AdaBoost algorithm *Ada-Stump* for the remainder of this paper. Fig. (1.b) displays a failure example of the Ada-Stump. We see that it can not deal with the ‘xor’ patterns, even with 100 stumps.

One solution to this problem is to adopt more powerful weak classifiers, such as decision tree, to the boosting algorithm. It was proposed by several authors [11, 18] and we call it Ada-Tree here for notional convenience (it is different from the AdaTree method [13]). However, using deci-

sion tree greatly increases the time and computational complexity of the boosting algorithm. Many vision applications were trained on very large datasets with each sample having thousands or even millions features [22]. This limits the use of decision tree or CART, and Ada-Stump remains mostly used in vision [22]. In this paper, we show that Ada-Stump intrinsically can not deal with the ‘xor’ problem. We propose layered logic models for classification, namely *Ada-Or*, *Ada-And*, and *Ada-AndOr*. The algorithm has several interesting properties: (1) it naturally incorporates the ‘and’, ‘or’, and ‘not’ relations in the algorithm; (2) it has much more discrimination power than Ada-Stump; (3) it has much smaller computational complexity than tree based AdaBoost with only slightly degraded classification performance.

A recent effort to combine ‘and’ and ‘or’ in AdaBoost has been proposed in [8]. However, the ‘and’ and ‘or’ relations are not naturally embedded in the algorithm and it requires very complex optimization procedure in training. How the algorithm can be used for general tasks in machine learning and computer vision is at best unclear.

We apply the proposed models, Ada-Or, Ada-And, and Ada-AndOr, on several typical datasets from the Irvine repository and two challenging vision applications, object segmentation and pedestrian detection. Among the models, Ada-AndOr performs the best nearly in all cases. We observe significant improvements on all the datasets over Ada-Stump. For pedestrian detection, the performance of Ada-AndOr is very close to HOG [7] using simple Haar features, though the main objective of this paper is not to develop a pedestrian detector.

2. AdaBoost algorithm

In this section, we briefly review the AdaBoost algorithm and explain why Ada-Stump fails on the ‘xor’ problem.

2.1. Algorithms and theory

Let $\{(x_i, y_i, D_1(i)), i = 1 \dots N\}$ be a set of training samples and $D_1(i)$ is the distribution for each sample x_i . AdaBoost algorithm [12] proposed by Freund and Schapire learns a strong classifier $H(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x))$, based on the training set, by sequentially combining a number of weak classifiers. We briefly give the general AdaBoost algorithm [12] below:

The AdaBoost algorithm minimizes the total error $\sum_i e^{-\sum_{t=1}^T \alpha_t y_i h_t(x_i)}$ by sequentially selecting h_t and computing α_t in a greedy manner. At each step, it is to minimize

$$E_t = \sum_i D_t(i) e^{-\alpha_t y_i h_t(x_i)}$$

by coordinate descent:

Given: $(x_1, y_1, D_1(1)), \dots, (x_N, y_N, D_1(N)); y_i \in \{-1, 1\}$
 For $t = 1, \dots, T$:

- Train weak classifier using distribution D_t .
- Get weak hypothesis $h_t : \mathcal{X} \rightarrow \{-1, +1\}$.
- Calculate the error of $h_t : \epsilon_t = \sum_{i=1}^N D_t(i) \mathbf{1}_{(y_i \neq h_t(x_i))}$.
- Compute $\alpha_t = -\log \epsilon_t / (1 - \epsilon_t)$.
- Update: $D_{t+1}(i) \leftarrow D_t(i) \cdot \exp(-\alpha_t y_i h_t(x_i))$ with $\sum_i D_{t+1}(i) = 1$.

Output the the strong classifier: $H(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x))$.

Figure 2. Discrete AdaBoost algorithm. $\mathbf{1}$ is an indicator function. The variations to the AdaBoost algorithms such as arc-gv [5], RealBoost and GentleBoost [11] differ mostly from the way h_t and α_t are computed.

(1) Select the best weak classifier from the candidate pool which minimizes E_t .

(2) Compute α_t by taking $\frac{dE_t}{d\alpha_t} = 0$, which yields

$$\alpha_t = -\log \epsilon_t / (1 - \epsilon_t).$$

A very important property of AdaBoost is that after a certain number rounds, the test error still goes down even the training error is not improving [4]. This makes AdaBoost less prone to the overfitting problem than many other classifiers. Schapire et al. [19] explained this behavior of AdaBoost from the margin theory perspective. For any data (x, y) ,

$$\text{margin}(x, y) = \frac{y \sum_t \alpha_t h_t(x)}{\sum_t \alpha_t}.$$

$\text{margin}(x, y)$ essentially gives the confidence of the estimation y to x . For any given θ , the overall test error is bounded by

$$\text{error}_{\text{test}}(x, y) \leq p[\text{margin}(x, y) \leq \theta] + \tilde{O}\left(\sqrt{\frac{d}{m\theta^2}}\right), \quad (2)$$

where d is the VC dimension of the weak classifier and m is the number of training samples. Eqn. (2) shows three directions to reduce the test error: (1) increase the margin (related to training error but not exactly the same); (2) reduce the complexity of the weak classifier; (3) increase the size of training data.

Moreover, it is shown [11] that AdaBoost and its variations are asymptotically approaching the posterior distribution (there are still some debates about this probabilistic formulation of the AdaBoost algorithm).

$$p(y|x) = \frac{e^{2y \sum_t \alpha_t h_t(x)}}{1 + e^{2y \sum_t \alpha_t h_t(x)}}. \quad (3)$$

The margin is directly tied to the discriminative probability.

2.2. The xor problem

It is well-known that the points shown in Fig. (1.a) as ‘xor’ are not linearly separable. The red and blue points are the positive and negative samples respectively. Each weak classifier makes a decision whether a point lies above or below a line passing the origin. Using this type of weak classifier, the AdaBoost algorithm is not able to separate the red points from the blue ones. It is easy to verify. For any positive sample (x_1, x_2) with $H(x_1, x_2) = \sum_{t=1}^T \alpha_t h_t(x_1, x_2) > 0$, then $(-x_1, -x_2)$ is a positive sample also. However

$$h_t(-x_1, -x_2) = -h_t(x_1, x_2), \forall h_t$$

and therefore $H(-x_1, -x_2) < 0$.

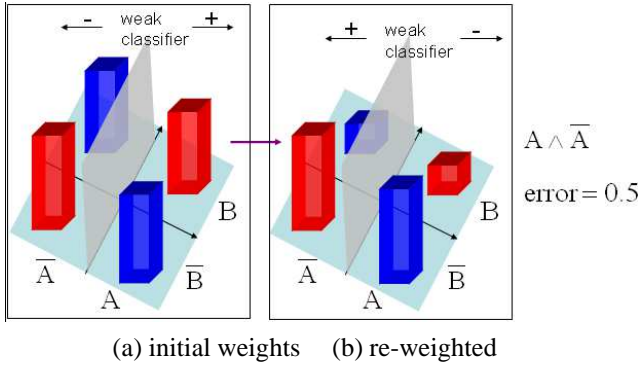


Figure 3. The re-weighting scheme in the AdaBoost to cause a deadlock.

Let

$$A = \{(x_1, x_2), x_1 > 0\} \text{ and } \bar{A} = \{(x_1, x_2), x_1 < 0\}, \text{ and}$$

$$B = \{(x_1, x_2), x_2 > 0\} \text{ and } \bar{B} = \{(x_1, x_2), x_2 < 0\}.$$

We denote \wedge, \vee, \neg as the ‘and’, ‘or’, and ‘not’ operations respectively. Thus, the positive samples in Fig. (1.a) can be denoted by

$$(A \wedge B) \vee (\bar{A} \wedge \bar{B}), \text{ or } (A \vee \bar{B}) \wedge (\bar{A} \vee B).$$

One of the key properties in AdaBoost is that it re-weights the training samples after each round by giving higher weights to those which were not correctly classified by the previous weak classifiers. We take a close look at the re-weighting scheme to the points in Fig. (1.a). Initially, all the samples receive equal weights, shown in Fig. (3.a). For any weak classifier (line passing the origin), the error is $\epsilon = 0.5$ which means that they are equally bad. In a computer simulation the value is usually slightly smaller than 0.5 since the training points are discretized samples. Once a weak classifier is selected, e.g., the line $x_1 > 0$ (A), then positive samples ($A \wedge B$) and negative samples ($\bar{A} \wedge B$)

are correctly classified, and they will receive lower weights. Fig. (3.b) shows the weights for the samples after the first step of the AdaBoost. Clearly, the weak classifier to minimize the error for this round would be $x_1 < 0$ (\bar{A}), which is a contradictory decision to the previous weak classifier (A). The re-weighted points after this round essentially lead the situation back to Fig. (3.a). The combination of the two weak classifiers is $A \wedge \bar{A} = \phi$ where ϕ denotes an empty set. The algorithm then keeps repeating the same procedure, which is a deadlock. Due to this reason, AdaBoost is sensitive to outliers since it keeps giving high weights to those miss-classified samples.

2.3. Possible solutions

The previous section shows that Ada-Stump cannot solve the ‘xor’ problem (on the line features passing the origin). The AdaBoost algorithm makes an overall decision based on a weighted sum $H(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x))$. It weakly performs the ‘and’, and ‘or’ operations on the weak classifiers. The ‘not’ is often embedded in the stump classifier by switching $>$ and $<$. We assume that all types of weak classifiers have the aspect of ‘not’ and we focus on ‘and’, and ‘or’ operations for the rest of this paper.

There are several possible ways to improve the algorithm:

1. Designing hyper features to allow the patterns to be linearly separable. For example, in the ‘xor’ case, it could be $x_1 \times x_2$. However, (1) it is often very hard to find the meaningful features which will nicely separate the positive and negative samples; (2) complex features often lead to the over-fitting problem.
2. Introducing the explicit ‘and’ and ‘or’ relations into the AdaBoost.

We can put ‘and’s on top of ‘or’s, or vice versa, or completely mix the two together. The probabilistic boosting tree (PBT) algorithm [20] is one way of recursively combining ‘and’s with ‘or’s. The disadvantages of PBT however are: (1) it requires longer training time than cascade and, (2) it produces complex classifier and may lead to overfitting (like the decision tree). Another solution is to build weak classifiers with embedded ‘and’ and ‘or’ operations. Using decision tree [16] as weak classifiers has been described in several papers [11, 18]. However, each tree is a complex classifier and it requires much longer time in training than the stump classifier. Also, it has more algorithm complexity than decision stump.

3. Layered logic classifiers

Eqn. (3) shows that the AdaBoost algorithm is essentially approaching a logistic probability by

$$p(y|x) \propto e^{y \sum_t \alpha_t h_t(x)} \propto \prod_t e^{y \alpha_t h_t(x)}.$$

The overall discriminative probability is a product of the probability of each h_t . Depending upon its weight α_t , each h_t makes a direct impact on $p(y|x)$. Using decision tree requires much longer time than stump classifier. This is particularly a problem in vision as we often face millions of image samples with each sample having thousands features.

Instead of using one layer AdaBoost, we can think of using two-layer AdaBoost with the weak classifier being stronger than decision stump, but simpler than decision tree. One idea might be to use Ada-Stump as weak classifier for the AdaBoost again, which we call Ada-Ada-Stump, or Ada-Ada for short notation. However, Ada-Ada still somewhat performs a linear summation and has difficulty on the ‘xor’ as well. Fig. (1.c) shows the positives classified by Ada-Ada with 50 weak classifiers of Ada-Stump, which itself has 5 stump weak classifiers. It is a failure example. It is worth to mention that one can indeed to make Ada-Ada work on these points by using very tricky strategies of randomly selecting a subset of points in training. However this greatly increases the training complexity and the procedures are not general.

Our solution is to propose AndBoost and OrBoost algorithms in which the ‘and’ and ‘or’ operations are explicitly engaged. We give detailed descriptions below.

3.1. OrBoost

For a combined classifier, we can use the ‘or’ operation directly by

$$H(x) = \text{sign}(h_1(x) \vee \dots \vee h_T(x)), \quad (4)$$

where

$$h_i(x) \vee h_j(x) = \begin{cases} +1 & \text{if } h_i(x) = +1 \text{ or } h_j(x) = +1 \\ -1 & \text{otherwise} \end{cases} \quad (5)$$

Fig (4) gives the detailed procedure of the OrBoost algorithm, which is straight forward to implement. The overall classifier is a set of ‘or’ operations on weak classifier, e.g. decision stump, and it favors positive answer. If any weak classifier provides a positive answer, then the final decision is positive, regardless of what other weak classifier will say. Unlike in the AdaBoost algorithm where mis-classified samples are given higher weights in the next round, OrBoost gives up some samples quickly and focus on those which can be classified correctly. This helps to solve the deadlock situation in AdaBoost shown in Fig. (3).

Given: $(x_1, y_1, D(1)), \dots, (x_N, y_N, D(N)); y_i \in \{-1, 1\}$
 For $t = 1, \dots, T$:

- Train weak classifier using distribution D .
- Get weak hypothesis $h_t : \mathcal{X} \rightarrow \{-1, +1\}$.
- Train weak classifier using weights D to minimize error $\sum_i D(i) \mathbf{1}(y_i \neq (h_1(x_i) \vee \dots \vee h_t(x_i)))$.
- The algorithm also stops if the error is not decreasing.

Output the overall classifier: $H(x) = \text{sign}(h_1(x) \vee \dots \vee h_T(x))$.

Figure 4. OrBoost algorithm.

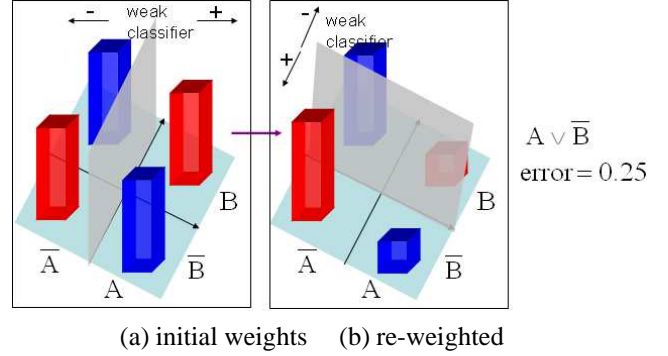


Figure 5. The re-weighting scheme in the OrBoost algorithm which breaks the deadlock in the AdaBoost algorithm.

Fig. (5) shows the feature selection and re-weighting steps by the OrBoost algorithm for the xor problem. The first weak classifier is selected the same as before ($x_1 > 0 = A$). However, positives AB and negatives $\bar{A}\bar{B}$ receive low weights in the AdaBoost since they have been classified correctly. This creates a deadlock. In OrBoost, the situation is different. Note that although the weights for all the samples D are fixed, the error evaluation function $\sum_i D(i) \mathbf{1}(y_i \neq (h_1(x_i) \vee \dots \vee h_t(x_i)))$ affects how $D(i)$ plays a role. This is similar to the re-weighting scheme in the AdaBoost. For example, positives AB and negatives $\bar{A}\bar{B}$ have been classified as positives by the first weak classifier, $x_1 > 0 = A$. The errors on AB and $\bar{A}\bar{B}$ are therefore decided already regardless what the later weak classifiers will be. Therefore, the second weak classifier would be $x_2 < 0 = \bar{B}$. The total error by the two combined weak classifiers is 0.25.

3.2. AndBoost

If we swap the labels of the positives and negatives in training, the ‘or’ operations in OrBoost can be directly turned into ‘and’ operations since

$$A \wedge B = \bar{A} \vee \bar{B}.$$

However, for a given set of the training samples, ‘and’ operations may provide complementary decisions to the ‘or’

operations. Similarly, we can use the ‘and’ operation directly by

$$H(x) = \text{sign}(h_1(x) \wedge \dots \wedge h_T(x)), \quad (6)$$

where

$$h_i(x) \wedge h_j(x) = \begin{cases} +1 & \text{if } h_i(x) = +1 \text{ and } h_j(x) = +1 \\ -1 & \text{otherwise} \end{cases} \quad (7)$$

Therefore, we can design an AndBoost algorithm in Fig. 6 which is very similar to the OrBoost algorithm in Fig. 4.

Given: $(x_1, y_1, D(1)), \dots, (x_N, y_N, D(N)); y_i \in \{-1, 1\}$
 For $t = 1, \dots, T$:

- Train weak classifier using distribution D .
- Get weak hypothesis $h_t : \mathcal{X} \rightarrow \{-1, +1\}$.
- Train weak classifier using weights D to minimize error $\sum_i D(i) \mathbf{1}(y_i \neq (h_1(x_i) \wedge \dots \wedge h_t(x_i)))$.
- The algorithm also stops if the error is not decreasing.

Output the overall classifier: $H(x) = \text{sign}(h_1(x) \wedge \dots \wedge h_T(x))$.

Figure 6. AndBoost algorithm.

The performance of the AndBoost on the ‘xor’ problem is the same as the OrBoost algorithm.

3.3. AdaOrBoost

After the introduction of the OrBoost and AndBoost algorithms, we are ready to discuss the proposed layered models. We simply use a two-layer AdaBoost algorithm with the weak classifiers in the second layer being the choice of OrBoost, AndBoost, or both. We call the models, *Ada-Or*, *Ada-And*, and *Ada-AndOr* respectively.

There are two levels of weak classifiers now. For Ada-Or, the OrBoost is its weak classifier. For OrBoost, any type of classifier can be its weak classifier. To keep the complexity of OrBoost and AndBoost under check, we simply use the decision stump. As we mentioned before the ‘not’ operation is naturally embedded in the decision stump. Therefore, the Ada-AndOr has all the aspects of logic operations, ‘and’, ‘or’, and ‘not’. Again, we call the weak classifiers in OrBoost and AndBoost *operations* to avoid confusion.

Fig. (1.b) shows the points which are classified by Ada-Stump using 100 stump weak classifiers. This failure example verifies our earlier claim for the ‘xor’ pattern. Fig. (1.d) shows the result by Ada-Or using 10 OrBoost weak classifiers, in which there are 2 or operations. As we can see, the positive samples have been classified correctly. Fig. (1.c) gives the result by Ada-Ada.

The margin theory of Ada-Or, Ada-And, and Ada-AndOr still follows the same as pointed by Schapire et al. [19] in eqn. (2). The complexity d of weak classifier

is decided by the OrBoost and AndBoost algorithms, which are just a sequence of ‘or’ operations or ‘and’ operations. It is slightly more complex than decision stump, but much simpler than decision tree or CART. It is worth to mention that both OrBoost and AndBoost include a special case where only one operation presents. This happens when the training error is not improving by adding the second operation. Therefore, stump classifier is also included in OrBoost and AndBoost, if stump is the choice of operator.

3.4. Experiments

There are several major issues we are concerned with for the choice of different classifiers for applications in machine learning and computer vision.

1. *Classification power*: This is often referred to as training error or margin in eqn. (2). A desirable classifier should produce low error and large margin on the training data.
2. *Low complexity*: This is often called VC dimension [21] and a classifier with small VC dimension often has a good generalization power, small difference between the training error and test error.
3. *Size of training data*: In the VC dimension and margin theory, the overall test error is also greatly decided by the availability of training data. The more training data we have and the classifier can handle, the smaller difference is between training error and test error. In reality, we often do not have enough training data since collecting them is not a easy task. Also, some none-parametric classifiers can only deal with limited amount of training data since they work on the kernel space, which explodes on large size data.
4. *Efficient training time*: For many applications in computer vision and data mining, the training data size can be immense and each data sample also has a large number of features. This demands an efficient classifier in training also. Fast training is more required in online learning algorithms [15] which has recently received many attentions in tracking.
5. *Efficient test time*: Judging the performance of a classifier is ultimately done in the test stage. A classifier is expected to be able to quickly give an answer. For many modern classifiers, this is not particularly a problem.

The first three criterion collectively decide the test error of a classifier. Another major factor affecting the performance a classifier is feature design. If the intrinsic features can be found, different types of classifiers will probably have a similar performance. However, the discussion of feature

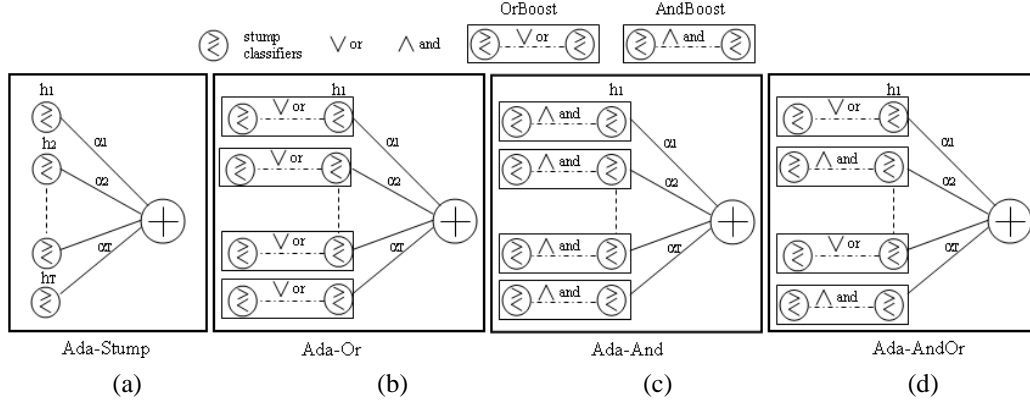


Figure 7. Two-layer logic models. (a) shows a traditional Ada-Stump algorithm. (b) displays the Ada-Or algorithm with OrBoost being the choice of weak classifier in the second layer. (c) and (d) give the illustrations of Ada-And and Ada-AndOr respectively.

design is out of the scope of this paper. Next, we focus on the performance of AdaOrBoost with comparison to the other classifiers.

3.5. Results on UCI repository datasets

One of the reasons that the AdaBoost algorithm is widely used is due to nice generalization power. Schapire et al. gave an explanation based on the margin theory after Breiman [4] observed an interesting behavior of AdaBoost: the test error of AdaBoost further asymptotically goes down even the training error is not decreasing. This was explained in the margin theory as to increase the margin with more weak classifiers combined. Breiman [5] then designed an algorithm called ‘arc-gv’ which tries to directly maximize the minimum margin in computing the α_t for AdaBoost. The experimental results were however contradictory to the theory since arc-gv produces bigger test error than AdaBoost. Reyzin and Schapire [18] tried to explain this finding and showed that the bigger test error by arc-gv was indeed due to the use of complex weak classifier, CART. Next we compare Ada-Or, Ada-And, and Ada-AndOr with arc-gv and AdaBoost using CART and decision stump.

We use the same datasets shown in Reyzin and Schapire [18], which are all from the UCI repository: breast cancer, ionosphere, ocr49 and splice. The datasets have been slightly modified the same way as in [18]. The two splice categories were merged into one in the splice dataset to create two-class data. Only digits 4 and 9 from the NIST database were used in the ocr49 dataset. The cancer, ion, ocr49 and splice then have 699, 351, 6000, 3175 data points respectively. Each sample usually has 20 – 60 features, depending upon what dataset it belongs to. The data samples are randomly split into training and testing for 10 trials. Table 1 shows the corresponding numbers.

To illustrate the effectiveness of the layered models, we first compare its results to those by Ada-Stump. Though

	cancer	ion	ocr 49	splice
training	630	315	1000	1000
test	69	36	5000	2175

Table 1. The sizes of training and test data of four datasets from UCI repository. The training and test data samples are randomly selected.

there are other alternatives such as RealBoost and GentleBoost [11], decision stump remains being widely adopted in the AdaBoost implementation. Fig. (8.a) shows the training and test errors on the splice dataset by Ada-Stump, Ada-Or, Ada-And, and Ada-AndOr using different number of weak classifiers. In the implementation of OrBoost and AndBoost, we use 5 ‘or’ operations. Each curve is averaged over 10 trials by randomly selecting 1000 samples for training and 2175 samples for testing. The Ada-AndOr gives the best performance among all. We also observe that the differences between the training and test errors for Ada-Stump and others are very similar. The results for real-world vision applications also show similar behavior of Ada-Or, Ada-And, and Ada-AndOr. This suggests that the OrBoost and AndBoost algorithms are having similar generalization power as decision stump.

To show how the use of different number of operations is affecting the performance, we conduct another experiment on the splice dataset. We plot out the training and test errors by using 50 weak classifiers with varying number of operations. The overall performance of the models, both in training and testing, is not improving too much with more than 3 operations shown in Fig. (8.b). Similar observations apply to other datasets as well. This suggests that the significant improvement can be achieved without introducing too much overhead.

It has been suggested [11, 5, 18] that the best performance of boosting algorithm is achieved by AdaBoost using decision tree [16] or CART [2]. Some of the confusions about generalization (test) error based on the margin theory

	<i>arc-gv-CART</i>	<i>Ada-CART</i>	<i>Ada-Stump(2500)</i>	<i>Ada-Or</i>	<i>Ada-And</i>	<i>Ada-AndOr</i>
	<i>arc-gv-stump</i>	<i>Ada-Stump</i>	<i>Ada-Stump</i>	<i>Ada-Stump</i>	<i>Ada-Stump</i>	<i>Ada-stump</i>
breast cancer	73.3%	57.3%	80.7%	57.4%	48.2%	56.0%
ionosphere	74.7%	36.1%	136.5%	63.6%	87.8%	66.7%
ocr 49	37.3%	32.5%	91.4%	57.2%	54.8%	38.2%
splice	47.8%	46.8%	113.6%	83.8%	68.4%	60.8%

Table 2. Test error ratios on the UCI datasets by arc-gv-CART, Ada-CART, Ada-Or, Ada-And, and Ada-AndOr over Ada-Stump. 500 weak classifiers are used in all cases except for Ada-Stump (2500) and Ada-Stump, which use 2500 and 100 stumps respectively. Ada-Or, Ada-And, and Ada-AndOr all contain 5 operations in the OrBoost and AndBoost which have roughly 2500 stumps for each. Ada-AndOr significantly outperforms Ada-Stump, and it shows to be comparable to arg-gv-CART, and is only a bit worse than Ada-CART.

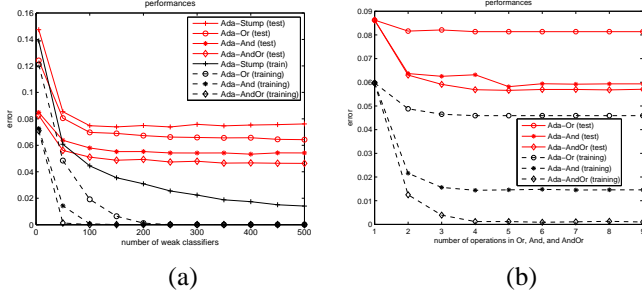


Figure 8. Training and test errors on the splice dataset by proposed models using different number of weak classifiers and operations. (a) displays comparison with different number of weak classifiers. Each curve is averaged over 10 trials by randomly splitting the dataset into training and test samples. (b) shows comparison of using different operations. Each algorithm uses 50 weak classifiers.

has recently been clarified by Reyzin and Schapire [18]. In table (2), we compare the our algorithms with AdaBoost and arc-gv using decision tree. For a fair comparison, we show the improvement of AdaOrBoost, arc-gv using CART, and AdaBoost using CART over those using decision stump. Table (2) shows the error ratio. As we can see, the improvement of AdaOrBoost is comparable to arc-gv using CART, but is worse than Ada-CART. However, each CART, after tree pruning, has around 16 leaf nodes with the tree depth being around 7. Therefore, the complexity of CART is much bigger than that of OrBoost and AndBoost. This is particularly an issue for applications in vision as the training data is massive with each data sample having thousands or even millions of features. The good performance of Ada-CART is achieved using an average of 7 levels of tree. This greatly limits its usage in many vision applications and leaves the decision stump classifiers still being currently widely used [22].

To illustrate the effectiveness of proposed algorithms, we further demonstrate them in two challenging vision problems, object segmentation and pedestrian detection.

First, we demonstrate it on the Weizmann horse dataset [3]. We use 328 images and use 126 for training and 214 for testing. Each input image comes with a label map in which the pixels on the horse body and background are labeled as 1 and 0 respectively. Given a test image, our task is

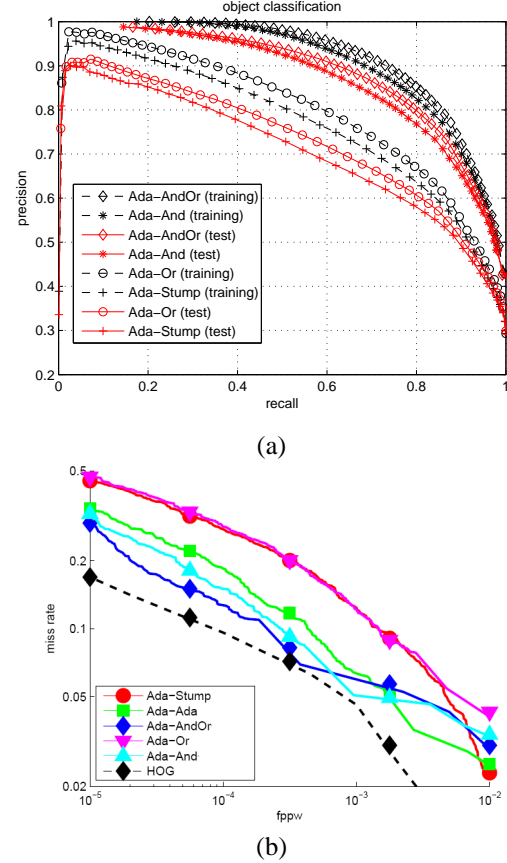


Figure 9. Precision and recall curves for the horse segmentation and error curve for pedestrian detection. (a) Shows the training and test errors by decision stump based Ada-Stump, Ada-Or, Ada-And, and Ada-AndOr. (b) displays the error curve by various classifiers. 3 operations are used for all classifiers. Ada-AndOr achieves the best result in both the cases among all Ada-. The horse segmentation result ($F=0.8$) outperforms that reported in [17] ($F=0.66$) and the pedestrian detection result is slightly worse than HOG [7].

to classify all the pixels into horse or background. In training, we take image patch of size 21×21 centered on every pixel as training samples. The background and horse body image patches are the negatives and positives respectively. For each image patch, we compute around 10,000 features such as the mean, variance, and Haar responses of the original as well as Gabor filtered. We implement a cascade

approach [22] and implement several versions. One uses Ada-Stump and others use Ada-Or and Ada-AndOr. Each cascade node selects and fuses 100 weak classifiers. All the algorithms use an identical set of features and bootstrapping procedure. Fig. (9.a) shows the precision and recall curve of the algorithms on the training and test images. We observe similar result as that for the UCI repository datasets. Ada-AndOr improves the results over Ada-Stump by a considerable amount. The differences between the training and test errors are nearly the same in this cascade setting as well. The F-value of the results by Ada-AndOr is around 0.8 which is better than the number 0.66 reported in [17] which uses low and middle level information.

Next, we show the Ada-AndOr algorithm for pedestrian detection on dataset reported in [7]. We use 8 level of cascade with different choices of weak classifiers for AdaBoost. Fig. (9b) shows the results by Ada-Stump, Ada-Ada, Ada-Or, Ada-And, and Ada-AndOr. The conclusion is nearly the same as before. Ada-AndOr achieves the best result among all with Ada-And being on the second place. Though we are not specifically addressing the pedestrian detection problem here, the result is nevertheless close to that by the well-known HOG pedestrian detector [7]. However, we only use a set of generic Haar features without tuning the system specifically for the pedestrian detection task.

3.6. Conclusions

Many of the classification problems in machine learning and computer vision can be understood as performing logic operations combining ‘and’, ‘or’, and ‘not’. In this paper, we have introduced layered logic classifiers. We show that AdaBoost can not solve the ‘xor’ problem using decision stump type of weak classifiers. We propose an OrBoost and AndBoost algorithms to study the ‘or’ and ‘and’ operations respectively. We demonstrate that the combined algorithm of two layers, Ada-AndOr, greatly outperformed Ada-Stump which is widely used in the literature. The improvement is significant in most the cases. We demonstrate the effectiveness of Ada-AndOr on traditional machine learning datasets, as well as challenging vision applications. Though decision tree based AdaBoost algorithm is shown to produce smaller test error, its complexity in training often limits its usage. The OrBoost and AndBoost algorithm only increases the time complexity slightly than decision stump, but they significantly reduce the test error. The Ada-AndOr algorithm is useful for a wide variety of applications in machine learning and computer vision.

Acknowledgment This work is supported by NSF IIS-1216528 (IIS-1360566) and NSF CAREER award IIS-0844566 (IIS-1360568).

References

- [1] C. M. Bishop, “Neural networks for pattern recognition”, *Oxford University Press*, 1995. 1
- [2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone CJ, “Classification and Regression Trees”, Chapman and Hall (Wadsworth, Inc.): New York, 1984. 1, 2, 6
- [3] E. Borenstein, E. Sharon and S. Ullman, “Combining top-down and bottom-up segmentation”, *Proc. IEEE workshop on Perc. Org. in Com. Vis.*, June 2004 7
- [4] L. Breiman, “Arcing classifiers”, *The Annals of Statistics*, 26, pp 801-849, 1998. 2, 6
- [5] L. Breiman, “Prediction games and arcing classifiers”, *Neural Computation* 11, 1493-1517, 1999. 1, 2, 6
- [6] R. Caruana and A. Niculescu-Mizil, “An Empirical Comparison of Supervised Learning Algorithms”, *Proc. of ICML*, 2006. 1
- [7] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *CVPR*, 2005. 2, 7, 8
- [8] M. Dundar and J. Bi, “Joint optimization of cascaded classifiers for computer aided detection”, *Proc. of CVPR*, 2007. 2
- [9] P. Dollár, Z. Tu, and S. Belongie, “Supervised learning of edges and object boundaries”, *Proc. of CVPR*, 2006.
- [10] R. O. Duda and P. E. Hart, “Pattern classification”, Wiley Interscience, 2000.
- [11] J. Friedman, T. Hastie and R. Tibshirani, “Additive logistic regression: a statistical view of boosting”, *Dept. of Stat., Stanford U. Te. Rep.* 1998. 1, 2, 3, 6
- [12] Y. Freund and R. E. Schapire, “A Decision-theoretic Generalization of On-line Learning And An Application to Boosting”, *J. of Comp. and Sys. Sci.*, 55(1), 1997. 1, 2
- [13] E. Grossmann, “AdaTree: Boosting a Weak Classifier into a Decision Tree”, *Proc. CVPR workshop on learning in computer vision and pattern recognition*, 2004. 1
- [14] D. Martin, C. Fowlkes, and J. Malik, “Learning to detect natural image boundaries using local brightness, color and texture cues”, *IEEE PAMI*, 26(5), 530-549, May 2004.
- [15] N. Oza and S. Russell, “Online Bagging and Boosting”, *Proc. of 8th International Workshop on Artificial Intelligence and Statistics*, 2001. 5

- [16] J.R. Quinlan, "Improved use of continuous attributes in C4.5", *J. of Art. Intell. Res.*, 4, pp. 77-90, 1996. 1, 3, 6
- [17] X. Ren, C. Fowlkes, and J. Malik, "Cue integration in figure/ground labeling", *Proc. of NIPS*, 2005. 7, 8
- [18] L. Reyzin and R. E. Schapire, "How boosting the margin can also boost classifier complexity", *Proc. of the 23rd International Conference on Machine Learning*, 2006. 1, 3, 6
- [19] R. E. Schapire, R. E. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics", 26, pp. 1651-1686, 1998. 2, 5
- [20] Z. Tu, "Probabilistic boosting tree: Learning discriminative models for classification, recognition, and clustering", *Proc. of ICCV*, 2005. 3
- [21] V. Vapnik, "Statistical Learning Theory". Wiley-Interscience, 1998. 5
- [22] P. Viola and M. Jones, "Robust Real-Time Face Detection", *Int'l J. of Comp. Vis.*, vol. 57, no. 2, pp. 137-154, 2004. 2, 7